RFC 9755 : IMAP Support for UTF-8

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 18 mars 2025

Date de publication du RFC : Mars 2025

https://www.bortzmeyer.org/9755.html

Successeur du RFC 6855 [1], voici la normalisation de la gestion d'Unicode par le protocole IMAP (dans sa version 4rev1) d'accès aux boîtes aux lettres.

Normalisé dans le RFC 3501, IMAP version 4rev1 permet d'accéder à des boîtes aux lettres situées sur un serveur distant. Ces boîtes peuvent avoir des noms en Unicode, les utilisateurs peuvent utiliser Unicode pour se nommer et les adresses gérées peuvent être en Unicode. L'encodage utilisé est UTF-8 (RFC 3629). C'est donc une des composantes d'un courrier électronique complètement international (RFC 6530). (Il existe un RFC équivalent pour POP le RFC 6856. Pour IMAP 4rev2, normalisé dans le RFC 9051, UTF-8 est prévu dès le début.)

Tout commence par la possibilité d'indiquer le support d'UTF-8. Un serveur IMAP, à l'ouverture de la connexion, indique les extensions d'IMAP qu'il gère et notre RFC normalise `UTF8=ACCEPT` (section 3). En l'utilisant, le serveur proclame qu'il sait faire de l'UTF-8. Par le biais de l'extension `ENABLE` (RFC 5161), le client peut à son tour indiquer qu'il va utiliser UTF-8. Une fois que cela sera fait, client et serveur pourront faire de l'IMAP UTF-8. Cette section 3 détaille la représentation des chaînes de caractères UTF-8 sur le réseau.

Les nouvelles capacités sont toutes décrites dans la section 10 et enregistrées dans le registre IANA <https://www.iana.org/assignments/imap-capabilities/imap-capabilities.xhtml#imap-capabilities-1>.

On peut donc avoir des boîtes aux lettres qui ne peuvent être manipulées qu'en UTF-8. Les noms de ces boîtes doivent se limiter au profil « Net-Unicode » décrit dans le RFC 5198, avec une restriction supplémentaire : pas de caractères de contrôle.

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc6855.txt

Il n'y a bien sûr pas que les boîtes, il y a aussi les noms d'utilisateurs qui peuvent être en Unicode, et la section 5 spécifie ce point. Elle note que le serveur IMAP s'appuie souvent sur un système d'authentification externe (comme `/etc/passwd` sur Unix) et que, de toute façon, ce système n'est pas forcément UTF-8. Prudence, donc.

Aujourd'hui, encore rares sont les serveurs IMAP qui gèrent l'UTF-8. Mais, dans le futur, on peut espérer que l'internationalisation devienne la norme et la limitation à US-ASCII l'exception. Pour cet avenir radieux, la section 7 du RFC prévoit une capacité `UTF8=ONLY`. Si le serveur l'annonce, cela indique qu'il ne gère plus l'ASCII seul, que tout est forcément en UTF-8.

Outre les noms des boîtes et ceux des utilisateurs, cette norme IMAP UTF-8 permet à un serveur de stocker et de distribuer des messages dont les en-têtes sont en UTF-8, comme le prévoit le RFC 6532.

La section 8 expose le problème général des clients IMAP très anciens. Un serveur peut savoir si le client accepte UTF-8 ou pas (par le biais de l'extension `ENABLE`) mais, si le client ne l'accepte pas, que peut faire le serveur ? Le courrier électronique étant asynchrone, l'expéditeur ne savait pas, au moment de l'envoi si tous les composants, côté réception, étaient bien UTF-8. Le RFC n'impose pas une solution unique. Le serveur peut dissimuler le message problématique au client archaïque. Il peut générer un message d'erreur. Ou il peut créer un substitut, un ersatz du message originel, en utilisant les algorithmes du RFC 6857 ou du RFC 6858. Ce ne sera pas parfait, loin de là. Par exemple, de tels messages « de repli » auront certainement des signatures invalides. Et, s'ils sont lus par des logiciels différents (ce qui est un des avantages d'IMAP), certains gérant l'Unicode et d'autres pas, l'utilisateur sera probablement très surpris de ne pas voir le même message, par exemple entre son client traditionnel et depuis son "*webmail*". C'est affreux, mais inévitable : bien des solutions ont été proposées, discutées et même décrites dans des RFC (RFC 5504) mais aucune d'idéale n'a été trouvée.

Je n'ai pas testé les implémentations en logiciel libre mais, apparemment, la bibliothèque standard <`https://docs.python.org/3/library/imaplib.html`> de Python est déjà conforme à ce RFC (cherchez "UTF" dans la documentation), ainsi que celle de Java. Si vous connaissez des utilisateurices du courrier électronique en Unicode, demandez leur quel serveur IMAP ielles utilisent.

Le résumé des différences (peu importantes) avec son prédécesseur, le RFC 6855 est dans l'annexe B. `UTF8=APPEND` disparait, et `FETCH BODYSTRUCTURE` voit sa sémantique modifiée.